0.1 Question 1: Unboxing the Data

0.1.1 Question 1a

As mentioned above, we are working with just one month of data. In the full database (which we don't have access to), tables like the `data` table have billions of rows. What do you notice about the design of the database schema above that helps support the large amount of data and minimize redundancy? Keep your response to at most three sentences.

Hint: There is no need to examine any data here. What is a technique learned in lecture 16? Define that technique.

The design pf the database schema above uses normalization. By definition, normalization is the process of splitting (decomposing) relations into multiple relations to minimize redundancy.

0.1.2 Question 1d

Do you see any issues with the schema given? In particular, please address the two questions below: - Can you uniquely determine the building given the sensor data? Why? (**Hint:** given a row in the `data` table, can you determine a **uniquely** associated row in `real_estate_metadata` table? Your answer should draw insights from 1b.) - Could `buildings_site_mapping.building` be a valid foreign key pointing to `real_estate_metadata.building_name`? (**Hint:** think about the definition / constraints of a foreign key.)

Please keep your response to **at most three sentences.**

We can't uniquely determine the building given the sensor data because the intermediate table buildings_site_mapping maps to multiple tuples from real_estate_metadata (from 1b) given a bulding value. In addition, buildings_site_mapping.building couldn't be a valid foreign key pointing to real_estate_metadata.building_name because it has to point to a UNIQUE real_estate_metadata.building_name (or in other word, real_estate_metadata.building_name has to be a primary key).

0.2 Question 3: Entity Resolution

0.2.1 Question 3a

There is a lot of mess in this dataset related to entity names. As a start, have a look at all of the distinct values in the `units` field of the `metadata` table. What do you notice about these values? Are there any duplicates? **Limit your response to one sentence.**

There are some duplicates in the units values (e.g. Gal vs gal vs Gallons).

0.2.2 Question 3d

Moving on, have a look at the `real_estate_metadata` table—starting with the distinct values in the `location` field! What do you notice about these values? Keep your response to at most two sentences.

There are some minor typos in the name of location (e.g. FRANCISC O instead of FRANCISCO).